Supplementary Online Document

Identifying Predictors for Energy Poverty in Europe Using Machine Learning

October 8, 2021

1 Data

Enable-EU provides the data set used for this research (Enable-EU, 2018). In this section the data set is described, and the preprocessing steps are discussed.

1.1 Data description

The data set is available on the website of Enable-EU and was published in March 2018. It is derived from a survey conducted in eleven countries in Europe. Income and energy costs, comprising of heating and electricity, are reported for households, making the data set eligible to apply the framework to. The survey, accompanied by instructions for the conducting parties, is published along with it. The data set is published in Excel format, containing 473 columns and 11.265 rows. The diverse group of eleven countries can be considered representative of Europe, and thus serves as an excellent sample to conduct our research of finding energy poverty predictors valid across Europe.

The survey is divided into several different sections: general questions, mobility, shift to prosuming, heating and cooling, use of electricity, and governance framework. Only the general questions were surveyed in each country, namely a section 'home / building characteristics and household possessions' and 'social and economic characteristics'. As the research aims to find European predictors, the data set was limited to only questions in these two sections.

The general section comprised of 24 questions, with sub questions, resulting in 86 columns being available. Most questions required respondents to answer categorically, the answers were encoded in the data set by integer values representing the categories. For each of the columns there was a description of the question it represented available in the Excel file, and the correspondence between the answers and the integer values was also provided. Missing values were handled inconsistently: for many columns there was no data, represented by a Not a Number (NaN) value, for some there was a special integer value indicating that no value was available, and for some there was an additional indicator column to indicate that the respondent did not answer the question.

1.2 Preprocessing

The data set had to go through several preprocessing steps before it was ready to be used. Data preprocessing steps and data handling were mostly performed using the Python programming language (van Rossum and Drake Jr, 1995), in combination with the pandas library (McKinney et al., 2010).

In order to label the households in our data set, two variables were necessary: income and energy expenditure. The former could either be reported monthly or annually, therefore a simple check whether one of these two held a valid value was sufficient to guarantee income was filled. Energy expenditure was derived from 7 columns regarding heating and electricity cost. Two of these were columns indicating whether the question was answered by a respondent, four concerned yearly and monthly, heating and electricity costs. The last column corresponded to a question on the number of months the households

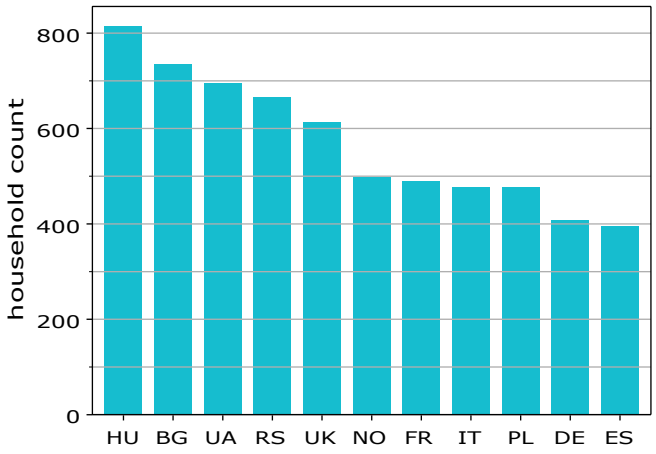

Figure 1: Number of respondents from each respective country in the data set after cleaning.

paid for heating in the last heating season. After this data cleaning, the number of respondents per country ranged between 711 and 1500 respondents and is plotted in Figure 1.

Income was reported in two columns depicting yearly and monthly income, corresponding to how the people in the respective country generally calculate their income. The survey required respondents to classify their income into the corresponding decile in their respective country, leaving the results to be categorized into 10 brackets, represented by integer values 1 to 10. For some rows, the entry contained a NaN value; however, a value of 98 and 99 corresponds to 'refused to answer' and 'do not know' respectively. The data was cleaned accordingly and a column *income bracket* was engineered representing a uniform income measure.

The feature energy expenditure is constructed by summing the heating and electricity costs of each household, and - if necessary - convert them from monthly to yearly costs. The conversion of electricity monthly costs to annual costs is simply done by multiplying it by 12. Heating costs reported as monthly costs, had an accompanying column with a question on the number of months in which heating was required. If this column was filled, the costs were multiplied by it in order to obtain annual costs. If it was not filled, the costs were multiplied by the median number reported in the country. Costs were reported in the currency used in the country in question. Therefore, the yearly energy expenditures were all converted to euros using the first available exchange rate to euros in 2018. This was retrieved from online currency conversion tool xe (Xe.com Inc, 2020). The currency in each of the countries, and the exchange rate are tabulated in Table 1.

Household size was was derived from 6 columns making a distinction between age and gender of members of the household. Summing up the household composition gives us a household size, irrespective of the gender and age.

As absolute income was not reported in the survey, relative energy expenditure was not directly available. The thresholds used to delimit the income brackets of each country in the survey corresponds to the national statistics for each country. To assign an absolute income to every income decile, the disposable income of the European Union Statistics on Income and Living Conditions survey was used (Eurostat, 2018), such that all approximations came from the same source. For the deciles, the 9 cutoffs points were reported. Linear inter- and extrapolation was used to approximate an income corresponding to the reported decile, depicted in Figure 2. No data was available for Ukraine, therefore an estimation was performed based on the relative GDP per capita compared to Serbia, using data from the World Bank (World Bank, 2020).

country	currency	exchange rate
Bulgaria	Bulgarian lev	0.511292
France	Euro	1
Germany	Euro	1
Hungary	Hungarian forint	0.003226
Italy	Euro	1
Norway	Norwegian krone	0.102712
Poland	Polish złoty	0.239406
Serbia	Serbian dinar	0.008446
Spain	Euro	1
Ukraine	Ukrainian hryvnia	0.029633
United Kingdom	Pound sterling	1.124859

Table 1: Table of countries in the data set with their currency and the first available exchange rate to euros in 2018.

After all preprocessing was completed, the data could be labeled. The class distribution of the data set can be found in Table 2. The data set is imbalanced, indicating that the data points are not evenly distributed per class. The data was split into a training set and a test set, with the test set containing 20% of the samples. This was done in a stratified fashion which keeps the class distribution the same for both sets.

Risk group	Number of households	
No risk	2774	44.3%
Income risk	2228	35.5%
Expenditure risk	841	13.4%
Double risk	425	6.8%

Table 2: The energy poverty risk label distribution of the data set

2 Energy poverty classification framework

In Table 3, the 2M and M/2 indicators are tabulated for all countries in the data set using an approximated income (derived as described in Section 1.2). Western European countries have a 2M indicator that is close to the 10% indicator previously used in the UK. However, Eastern European countries spend a substantially higher share of their income on energy. Therefore, a uniformly set threshold, such as the 10% one, appears unfit as a European indicator. Moreover, the table suggests that the heterogeneity of the data causes a large deviation in both relative and absolute energy expenditure. This suggests that this threshold should be determined for every country individually.

2.1 Employed framework

The energy poverty classification framework proposed by Dalla Longa et al. requires an income and an energy expenditure threshold to be set (Dalla Longa et al., 2021). In that paper, the energy expenditure threshold is set at the 80th quantile. This approach was adapted for our research, such that every participating country has a distinct energy expenditure threshold corresponding to the 80th quantile.

The data presents a uniform income metric with the deciles it was reported in. This allows for one threshold to be set for all countries. The minimum wages in 2018 for each of the countries are

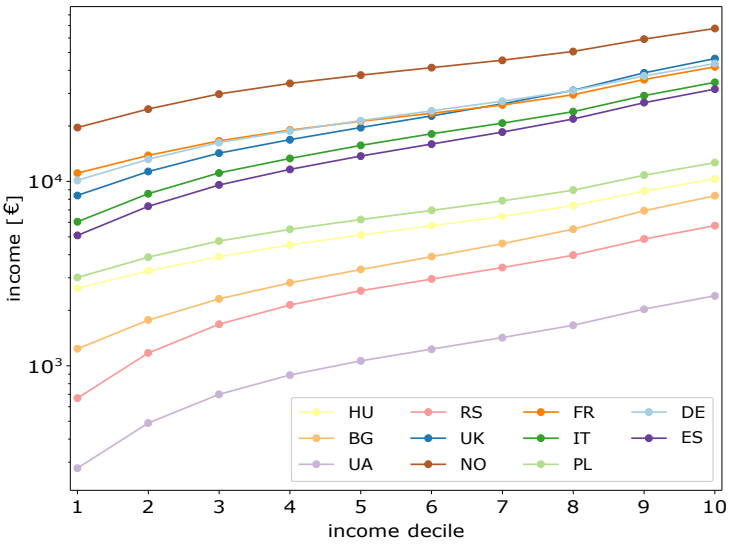

Figure 2: Approximated yearly incomes corresponding to each income deciles plotted for all countries.

retrieved from Eurostat and the corresponding deciles determined. The third income decile is the mode. Norway and Italy do not have a nationwide minimum wage (Rødseth and Holden, 1989; Tufo, 2018) and minimum wages for Ukraine were not available on Eurostat. A study conducted in Spain found that for three different energy poverty indicators, 99% of the households classified as being energy poor, are in the first three income deciles (Romero et al., 2018). This lead us to set the income threshold between the third and fourth deciles in the current research.

The thresholds used are relative: the 3rd decile of income and the 8th decile of energy expenditure. Consequently, every country has the same share of households that cross the thresholds and are thus at risk, irrespective of mean income or energy expenditure. This allows the framework to be applied to a heterogeneous set of countries such as ours. Due to the thresholds' statistical nature, if energy becomes cheaper, energy expenditure will decrease, and the corresponding threshold will move down with it. Similarly, if income increases, so will the income threshold. As a result, within this particular framework, policy measures targeting energy poverty eradication should aim at reducing the intersection between the datapoints above the expenditure threshold and those below the income threshold, i.e. the size of the double risk category. This category contains households that have the lowest incomes but the highest energy expenditures. The size of this risk group can be used as an indication of the prevalence of energy poverty in a set of households.

3 Model building

The data set available was imbalanced, as can be observed in Table 2. This can be circumvented by taking a subsample of the training data with the same number of points from each class, known as undersampling. Undersampling is a robust and effective approach to counter class imbalance (Liu et al., 2009). The data set used in this research does not have enough data points available to perform undersampling, as this would result in too few samples to effectively train a model. Another method is to subsample with replacement. This would include the same data point several times in the data set used by the model for training. This is known as oversampling. There are also advanced methods available to synthetically create additional data points of the minority class. Most methods use some

Country	Households	Median	2M	M/2	Minimum wage bracket
Norway	499	e 2311	12.6%	e 1156	-
United Kingdom	613	e 1215	13.3%	e 607	3
Germany	407	e 1900	13.7%	e 950	3
Spain	395	e 960	15.8%	e 480	3
France	489	e 1890	16.4%	e 945	3
Italy	478	e 1890	28.9%	e 945	-
Poland	477	e 862	37.5%	e 431	4
Hungary	815	e 890	48.2%	e 445	5
Bulgaria	734	e 798	53.1%	e 399	4
Ukraine	696	e 267	54.4%	e 133	-
Serbia	665	e 760	64.8%	e 380	7
Total	6268	e 997	31.1%	e 499	-

Table 3: The countries in the Enable data set with median annual energy expenditure, two conventional indicator values, and the income decile the minimum wage corresponds to.

variation of adding random points that lie on lines in feature space between two data points from the same class, to the training set (Hall et al., 2006; He et al., 2008).

As we are trying to understand the mechanism that drive energy poverty in Europe, artificially generated data points might lead to results not reflected by real data. Therefore, *class weights* were passed to the classifier. CatBoost multiplies the gradient of a sample with the weight corresponding to its true label. Weight were calculated on the data set relative to the reciprocal of a class count in the data set. This has a similar effect on training as oversampling. With oversampling, if from one class with two data points, one is doubly sampled. The gradient of this data point is used twice. Whereas, by using class weights, both data point's gradient is multiplied by 1.5. As a result, we get a 'smoother' form of oversampling.

The model is trained using weighted multiclass logistic regression loss function, depicted in Equation 1. The weights for each class are determined over the training set. Splitting the data that is stratified by their labels ensures that the training, validation, and test set have a similar class distribution.

$$-\frac{\sum_{i=1}^{N} w_i \log\left(\frac{e^{s_{i,y_i}}}{\sum_{j=0}^{M-1} e^{s_{i,j}}}\right)}{\sum_{i=1}^{N} w_i}, \tag{1}$$

Catboost has many tools to monitor training, all metrics specified by the user are plotted in real time. The model that performed best on the optimization metric, loss function, is saved. The library also enables the user to use a different metric to evaluate the performance of the model. This can be a non-differentiable function, that is evaluated at the end of each iteration. The evaluation metric is used to determine the best model, and the differentiable loss function is used for optimization, to calculate the gradient to construct the next tree for our ensemble.

For model evaluation, some important measures are the True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). For classification with more than one or two classes, all classes other than the one evaluated are assumed to be of the negative class. In a confusion matrix with absolute numbers, the TP representing the samples correctly classified as the class being evaluated is the value on the diagonal. TN is the number of samples correctly not classified as the class: all entries except the entries in the row and column corresponding to the class. FP are all samples that were incorrectly classified as the class: the sum of all entries in the column, except the diagonal. FN are all samples of our class that were not classified as being of the class: the sum of the entries in the row except the diagonal.

Precision $= \frac{TP}{TP+FP}$, and recall $= \frac{TP}{TP+FN}$ use the above described metrics to give one numberthat encapsulates the relevance of the found positives by the model. Both metrics can effortlessly be

maximized by either classifying all (recall), or only a minor fraction (precision) of the samples as positive. The F1 score is the harmonic mean of the precision and recall and evaluates a model on both metrics simultaneously. It is used as the evaluation metric in this study. It has long been known as the Sørensen–Dice coefficient to measure the degree of similarity between two sets (Sørensen, 1948; Dice, 1945) given in Equation 2.

$$F_1 = \frac{2|X \cap Y|}{|X|+|Y|} = \frac{2TP}{2TP+FN+FP} = \frac{2(Precision)*(Recall)}{Precision+Recall}, \quad (2)$$

However, since we have four classes, we calculate the F1 score for each of the classes and weigh them to get a total F1 score.

$$TotalF_1 = \frac{\Sigma_{k=0}^{M-1} w_k F1_k}{\Sigma_{k=0}^{M-1} w_k}, \quad (3)$$

The training and validation learning curves are plotted in Figure 3. The loss function starts around 1.3, which makes intuitive sense as then it would do little more than random guessing, which would result in a loss of $-log(\frac{1}{4}) \approx 1.39$ for all classes. Overfitting of the model can be observed in both plots where the training and validation curves start to diverge. The iteration with the best score on the validation set is indicated with a dashed black line. This is at iteration 126 with a TotalF1 score of 0.65, and at iteration 264 with a loss function value of 0.64. The model results generalized poorly at the point of minimum loss function value. We hypothesize this could be caused by some of the risk classes being so small that the model starts overfitting on the few samples that are in the training and validation set. A decision boundary tighter around these samples would decrease the loss but hinders the generalization of the model. Fivefold cross validation shows very similar learning curves with the best average TotalF1 score of 0.63 at iteration 118, iteration 254 minimizes the average loss function at a value of 0.65. These results confirm our findings and show that the results are robust.

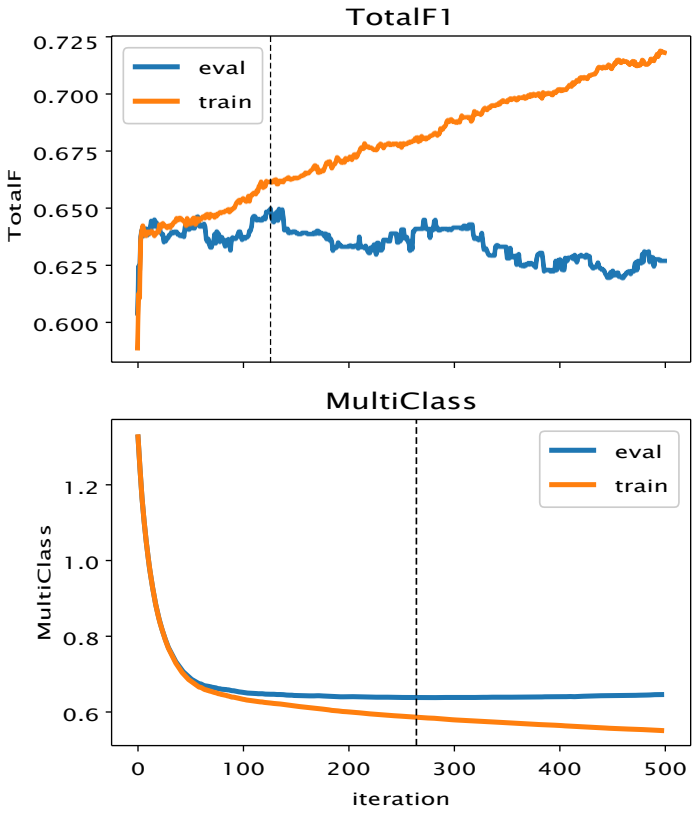

Figure 3: Scores on validation and training set during training time. The dashed black line indicates the iteration where the best validation score occurred.

References

F. Dalla Longa, B. Sweerts, and B. van der Zwaan. Exploring the complex origins of energy poverty in The Netherlands with machine learning. *Energy Policy*, 156:112373, 2021. ISSN 0301-4215. doi: https://doi.org/10.1016/j.enpol.2021.112373.

L. R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945. doi: 10.2307/1932409. URL https://doi.org/10.2307/1932409.

Enable-EU. D4.1 dataset for the comparative sociological analysis of the household survey results. http://www.enable-eu.com/wp-content/uploads/2019/10/enable_eu_dataset_households.zip, 2018. Accessed 01-04-2020.

Eurostat. Distribution of income by quantiles - eu-silc and echp surveys (ilc_di01), 2018. data retrieved from Eurostat 06-2020, https://ec.europa.eu/eurostat/web/products-datasets/-/ILC_DI01.

L. O. Hall, K. W. Bowyer, P. W. Kegelmeyer, and N. V. Chawla. SMOTE: Synthetic Minority Oversampling Technique Nitesh. *Journal of Artificial Intelligence Research*, 2009(Sept. 28):321–357, 2006. ISSN 10769757. doi: 10.1613/jair.953. URL https://arxiv.org/pdf/1106.1813.pdf{%}0Ahttp://www.snopes.com/horrors/insects/telamonia.asp.

H. He, Y. Bai, E. Garcia, and S. Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. *Proceedings of the International Joint Conference on Neural Networks*, pages 1322 – 1328, 2008. doi: 10.1109/IJCNN.2008.4633969.

X. Liu, J. Wu, and Z. Zhou. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2):539–550, 2009.

W. McKinney et al. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56. Austin, TX, 2010.

A. Rødseth and S. Holden. *Wage formation in Norway*. IIES, 1989.

J. C. Romero, P. Linares, and X. López. The policy implications of energy poverty indicators. *Energy Policy*, 115:98–108, 2018. ISSN 03014215. doi: 10.1016/j.enpol.2017.12.054. URL https://doi.org/10.1016/j.enpol.2017.12.054.

T. J. Sørensen. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. *Kongelige Danske Videnskabernes Selskab*, 5(4):1–34, 1948. URL https://ci.nii.ac.jp/naid/10008878962/en/.

M. Tufo. The minimum wage in italy during the eurozone crisis age and beyond. *IUSLabor. Revista d'anàlisi de Dret del Treball*, pages 205–231, 2018.

G. van Rossum and F. L. Drake Jr. *Python tutorial*. Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands, 1995.

World Bank. Gdp per capita, 2020. https://data.worldbank.org/indicator/NY.GDP.PCAP.CD?locations=UA-RS. Accessed April, 2020.

Xe.com Inc. Online foreign exchange tools and services company, 2020. www.xe.com. Accessed April, 2020.